The comparison of tree-sibling time consistent phylogenetic networks is graph isomorphism-complete

Gabriel Cardona[1], Mercè Llabrés [1], Francesc Rosselló [1], and Gabriel Valiente [2]

[1]Department of Mathematics and Computer Science, University of the Balearic Islands
07122 Palma de Mallorca, Spain

[1]Algorithms, Bioinformatics, Complexity and Formal Methods Research Group
Technical University of Catalonia, 08034 Barcelona, Spain

February 26, 2009

Abstract

In [2] we gave a metric on the class of semibinary tree-sibling time consistent phylogenetic networks that is computable in polynomial time; in particular, the problem of deciding if two networks of this kind are isomorphic is in P. In this paper, we show that if we remove the semibinarity condition above, then the problem becomes much harder. More precisely, we proof that the isomorphism problem for generic tree-sibling time consistent phylogenetic networks is polynomially equivalent to the graph isomorphism problem. Since the latter is believed to be neither in P nor NP-complete, the chances are that it is impossible to define a metric on the class of all tree-sibling time consistent phylogenetic networks that can be computed in polynomial time.

1 Introduction

After the realization that reticulation processes, like hybridizations, recombinations or lateral gene transfers, have been more relevant in the evolution of life on Earth than previously thought [6], there has been a growing interest in the development of algorithms for the reconstruction of *phylogenetic networks*: graphical models of evolutionary histories that go beyond phylogenetic trees by including *hybrid nodes* of in-degree greater than one representing reticulation events. As the number of available such algorithms increases, the need of methods for the comparison of phylogenetic networks also increases, as they are used, for instance, to assess the reliability and robustness of these algorithms [12, 14].

One of the types of phylogenetic networks for which there exist reconstruction methods [9, 10] are the *tree-sibling time consistent* networks, *TSTC networks*, for short (see §2 for a formal definition). There have been several at-

tempts to define a metric on the class of all TSTC networks on a given set of taxa [13], and we have recently given a metric on the class of all *semibinary* TSTC networks, where all hybrid nodes have in-degree two [2], but none of the metrics for phylogenetic networks computable in polynomial time proposed so far satisfies the separation axiom (distance 0 means isomorphism) for generic TSTC networks: see [3, 4]. In this paper we show why it should come as no surprise: such a metric would solve in polynomial time the graph isomorphism problem.

The graph isomorphism problem is one of the most important decision problems for which the computational complexity is not known yet [7, 11]. It is believed to be neither in P nor NP-complete, and subexponential time solutions for it are known. A problem is said to be *graph isomorphism-complete* when it is polynomially equivalent to the graph isomorphism problem. In this paper we show that, for every set S with more than two elements, the isomorphism problem for TSTC phylogenetic networks with taxa bijectively labeled in S is graph isomorphism-complete.

2 Preliminaries

Let $G = (V, E)$ be a non-empty rooted directed acyclic graph (a *rDAG*, for short). A node of G is a *leaf* if it has out-degree 0, *internal* if its out-degree is $\geqslant 1$, of *tree* type if its in-degree is $\leqslant 1$, of *hybrid* type if its in-degree is > 1, and *elementary* if it is a tree node of out-degree 1. A node v is a *child* of another node u (and, hence, u is a *parent* of v) if $(u, v) \in E$. Two nodes u and v are *siblings* of each other if they share a parent. An arc (u, v) in a rDAG is a *tree arc* when v is a tree node, and a *hybridization arc* when v is a hybrid node. The *height* of a node v is the longest length of a directed path from v to a leaf, and the *depth* of v is the longest length of a directed path from the root to v.

Given a finite set S of *labels*, a *S-rDAG* is a rDAG with its leaves injectively labelled by S. By an isomorphism of S-rDAGs we understand an isomorphism of directed graphs that preserves the labelling, that is, that maps each leaf in one network to the leaf with the same label in the other network (in particular, isomorphic S-rDAGs must have the same sets of actual leaf labels). In a S-rDAG, we shall always identify without any further reference every leaf with its label.

A *phylogenetic network* on a set S of *taxa* is a S-rDAG such that:

- No tree node is elementary.
- Every hybrid node has out-degree 1, and its single child is a tree node.

We will say that a phylogenetic network is *tree-sibling* if every hybrid node has at least one sibling that is a tree node.

A *temporal assignment* [1] on a network $N = (V, E)$ is mapping $\tau : V \to \mathbb{N}$ such that:

(a) If v is a hybrid node and $(u, v) \in E$, then $\tau(u) = \tau(v)$.

(b) If v is a tree node and $(u, v) \in E$, then $\tau(u) < \tau(v)$.

We will say that a phylogenetic network is *time-consistent* if it admits a temporal assignment. The following alternative characterization of time consistency will be used later. For a proof, see [1, 5].

Proposition 1 *Let $N = (V, E)$ be a phylogenetic network, let E_H be its set of hybridization arcs, and let $N^* = (V, E^*)$ be the directed graph with the same set V of nodes as N and set of arcs $E^* = E \cup \{(v, u) \mid (u, v) \in E_H\}$. Then, N is time consistent if, and only if, N^* does not have any cycle containing some tree arc of N.*

The underlying biological motivation for the definitions on phylogenetic networks introduced so far is the following. In a phylogenetic network, tree nodes model species (either extant, the leaves, or non-extant, the internal tree nodes), while hybrid nodes model reticulation events, where different species interact to create new species, the parents of the hybrid node being the species involved in this event and its single child being the resulting species. The tree children of a tree node represent direct descendants through mutation. The first condition in the definition of phylogenetic network says that every non-extant species is assumed to have at least two different direct descendants, be them by mutation or through some reticulation event. This is a very common restriction in any definition of phylogeny (be it a tree or a network), since species with only one child cannot be reconstructed from biological data.

The tree-sibling condition says then that, for every reticulation event, at least one of the species involved in it must have some descendant through mutation. This condition was introduced with the name *class I* in L. Nakhleh's PhD Thesis [13], and it has reappeared in several phylogenetic network reconstruction methods [9, 10]. As far as the time consistency goes, we understand that the time assigned to a node represents the time when the corresponding species existed, or when the reticulation event took place. The first condition in time consistency means then that the species involved in a reticulation event must coexist in time in order to interact, while the second condition means that speciation takes some amount of time to take place.

3 Main Results

It is well known [7, 15] that the isomorphism problem for rDAGs is graph isomorphism-complete. It turns out that the isomorphism problem for rDAGs with their leaves bijectively labeled in any given set of labels is also graph isomorphism-complete: since we have not been able to find a proof of this easy result in the literature, we provide one here.

Proposition 2 *For every non-empty set S of labels, the isomorphism for S-rDAGs is graph isomorphism-complete.*

Proof. Without any loss of generality, we assume that $S = \{1, \ldots, n\} \subseteq \mathbb{N}$.

Let us prove first that the isomorphism of S-rDAGs reduces to the isomorphism of rDAGs. For every S-rDAG G, let G' be the rDAG obtained from G by unlabelling its leaves and then, for each $k = 1, \ldots, n$, if G contained a leaf labeled with k, then adding to this leaf k tree-children leaves; see Fig. 1. The construction of G' from $G = (V, E)$ adds $O(n^2) \leqslant O(|V|^2)$ nodes and arcs, and therefore it is polynomial in the size of G. And G can be reconstructed from G' by simply replacing, for each $k = 1, \ldots, n$, the node of height 1 with k leaves by a leaf labeled with k. Then, it is straightforward to check that, for every pair of S-rDAGs G_1 and G_2 over S, $G_1 \cong G_2$ as S-rDAGs if, and only if, $G'_1 \cong G'_2$ as rDAGs.

Let us prove now that the isomorphism of rDAGs reduces to the isomorphism of S-rDAGs. For every rDAG G, let G'' be the S-rDAG obtained from G by adding a new node a, arcs from each leaf of G to a and finally adding one child leaf to a labeled 1; see Fig. 2. The construction of G'' from $G = (V, E)$ adds 2 nodes and $O(|V|)$ arcs, and therefore it is polynomial. And G can be reconstructed from G'' by simply removing its leaves and its only height 1 node, a, as well as all arcs pointing to a or to the leaves. It is straightforward to check that, for every pair of rDAGs G_1 and G_2 over S, $G_1 \cong G_2$ if, and only if, $G''_1 \cong G''_2$ as S-rDAGs.

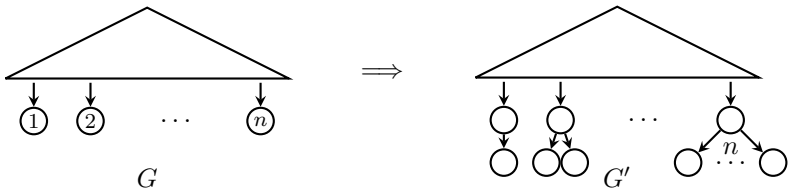

Figure 1: The construction involved in the reduction of the isomorphism of S-rDAGs to the isomorphism of rDAGs.

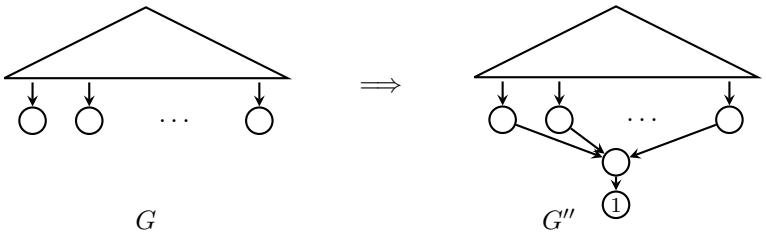

Figure 2: The construction involved in the reduction of the isomorphism of rDAGs to the isomorphism of S-rDAGs.

Theorem 1 *For every set S with $|S| \geqslant 3$, the isomorphism of TSTC-networks on a set S of taxa is graph isomorphism-complete.*

Proof. Without any loss of generality, we assume that $S = \{1, \ldots, n\} \subseteq \mathbb{N}$.

The isomorphism of TSTC-networks on S clearly reduces to the isomorphism of S-rDAGs, since the former are a special case of the latter. Let us prove now the converse reduction.

Let $N = (V, E)$ be a S-rDAG. Let $\overline{N}$ be the $(S \cup \{n+1, n+2\})$-rDAG obtained as follows:

(1) For every hybrid node h in N, remove all arcs from h to its children, and then add a new (tree) node u_h, an arc from h to u_h, and new arcs from u_h to the children of h in N. If h was a leaf, say with label k, then u_h becomes the new leaf labeled with k.

(2) For every hybridization arc $e = (v, h)$ in the resulting S-rDAG, split it into arcs (v, v_e) and (v_e, h), with v_e a new (tree) node.

Let N' denote the resulting S-rDAG after these two first steps.

(3) For every internal tree node v in N', add a new (tree) node v' and an arc (v, v').

(4) Split the arc (w, n) in N' pointing to the leaf n into two arcs (w, w_n) and (w_n, n).

(5) Add two new nodes a and b, and, for every node v' added in step (3), add arcs (v', a) and (v', b). Add also arcs (w_n, a) and (w_n, b). The nodes a and b will be hybrid.

(6) Add a tree leaf children labelled $n+1$ to a, and another one labelled $n+2$ to b.

An example of this construction is displayed in Fig. 3.

Let us prove now that $\overline{N}$ is a tree-sibling time consistent phylogenetic network.

- It is rooted (with the same root as N) and acyclic, because all new arcs are either used to split arcs in N into pairs of consecutive arcs, or to define paths that end in the new leaves $n+1$ or $n+2$ without forming cycles.
- It has no elementary nodes. Indeed, the internal tree nodes in N get an extra child in step (3), and the tree nodes that are added to N either get an extra child in step (3) or they get two children in (5).
- Its hybrid nodes have only one child, and it is a tree node: this is ensured for the hybrid nodes in N in step (1), and for the new hybrid nodes a and b by construction.

- It is tree-sibling. All hybrid nodes in N get a tree sibling in steps (2) and (3) (for every hybrid node h in N, if e is any arc pointing to h, then the tree child v'_e of the new node v_e added in the middle of e is such a tree sibling of h), and the hybrid nodes a and b have the tree sibling n.
- It is time consistent. To check this, we use Proposition 1 (and the notations introduced therein). Since we already know that $\overline{N}$ is acyclic, any cycle in $\overline{N}^*$ must contain some inverse of a hybridization arc. There are two possibilities for this inverse. If it has the form (h, x), with h one of the new hybrid nodes a or b introduced in step (5) and x one of the tree nodes v' introduced in step (3) or the tree node w_n introduced in step (4), then the only tree arcs that can be reached from x in $\overline{N}^*$ are those pointing to the leaves n, $n+1$ or $n+2$, and therefore no cycle in $\overline{N}^*$ contains this arc (h, x) together with a tree arc. And if this inverse is of the form (h, v_e), with h a hybrid node in N and v_e one of the tree nodes introduced in step (2), then it must be followed in the cycle by the arc (v_e, v'_e) added in step (3), and, as we have just said, the only tree arcs that can be reached from v'_e point to a leaf, and hence no cycle in $\overline{N}^*$ contains this arc (h, v'_e) and a tree arc, either.

It is clear that the construction of $\overline{N}$ from N adds $O(|V| + |E|)$ nodes and arcs to N, and thus it is polynomial in the size of N.

Now, the S-rDAG N can be easily reproduced from $\overline{N}$ by simply undoing its construction:

(5) Remove the leaves $n+1$ and $n+2$ and its hybrid parents a and b, together with all arcs pointing to them.

(4) Remove the elementary parent of the leaf n (which will be the remaining leaf with largest label in S) and replace it by an arc from the parent of the removed node to n.

(3) Remove all non-labeled leaves of the resulting rDAG together with the arcs pointing to them

(2) Remove each parent v_e of every hybrid node, and replace it by an arc from the parent of v_e to the hybrid child of v_e.

(1) Remove the only tree child of each hybrid node, and replace it by an arc from the hybrid node to each one of the children of the removed node.

(0) The resulting S-rDAG is N.

It is straightforward to check now that, for every pair of S-rDAGs N_1 and N_2, $N_1 \cong N_2$ if, and only if, $\overline{N_1} \cong \overline{N_2}$ as phylogenetic networks over $S \cup \{n+1, n+2\}$.

We cannot remove the condition $|S| \geqslant 3$ in the previous result because there are only two TSTC phylogenetic networks with less than 3 leaves (up to the actual names of the labels). In particular, this implies that, in the proof of the previous result, we cannot add less than 2 new leaves in the construction of $\overline{N}$ from N.

Proposition 3 *There is only one TSTC phylogenetic network on* $\{1\}$*, and only one TSTC phylogenetic network on* $\{1,2\}$*, and in both cases they are trees.*

Proof. The $\{1\}$-rDAG consisting of a single node, labeled 1, and the $\{1,2\}$-rDAG consisting of the phylogenetic tree with Newick code `(1,2);` are clearly TSTC phylogenetic networks. Let us check now that any other TSTC phylogenetic network has at least 3 leaves.

Let $N = (V, E)$ be a TSTC phylogenetic network other than those described in the last paragraph, let $\tau : V \to \mathbb{N}$ be a time assignment, and let v be an internal node with largest τ-value and, among those with this largest time assignment, of largest depth.

If v is a tree node, then all its children are either leaves or hybrid nodes with leaf children (because any tree descendant node of v has time assignment larger than $\tau(v)$). And v's hybrid children would have the same time assignment as v, but depth largest than v's depth, against the assumption. Therefore all children of v are leaves, and it has at least 2 children, because it cannot be elementary. Now, if v has more than 2 children, we are done, while if it has only two children, say the leaves 1 and 2, then v will have a parent in N (because N is not the tree `(1,2);`). If the parent of v is a tree node, let w be this node, and let z be another child of w. Since N does not contain cycles, and any path to 1 or 2 must contain w, we deduce that any descendant leaf of z must be different from 1 or 2: this gives at least 3 leaves. If, on the contrary, the parent of v is a hybrid node x, let w be the parent of x that has a tree child, say z. The time consistency prevents x to be a descendant of z (because $\tau(z) > \tau(w) = \tau(x)$) and therefore, since any path leading to 1 or 2 must contain x, any leaf that is a descendant of z will be different from $1, 2$: this gives again at least 3 leaves.

If v is a hybrid node, then its child is a leaf, say 1. Let v_1 be a parent of v that has a tree child. Since $\tau(v_1) = \tau(v)$ is the largest τ value of an internal node of N, this tree child must be a leaf, say 2. Now let v_2 be another parent of v. Since it is a tree node, it must have another child other than v, say x. If x is a tree node, it is a leaf, as we have just seen. If x is hybrid, then since $\tau(x) = \tau(v_2) = \tau(v)$, the tree child of x must be a leaf. In both cases, we obtain a leaf that is different from 1 and 2, that is, N contains at least 3 leaves.

4 Conclusion

We have proved that, unless the graph isomorphism problem belongs to P, there is no hope of defining a polynomially computable metric on the class of all TSTC phylogenetic networks on a set S of at least 3 taxa. It remains open the problem of defining polynomially computable, and biologically sound, metrics on the class of all TSTC phylogenetic networks on a given set S with all their hybrid nodes with in-degree bounded by some $d \in \mathbb{N}$. When $d = 2$, the μ-distance is such a metric [2], but it is no longer a metric for $d = 4$ (see the Supplementary Material to the aforementioned paper). Actually, we do not even know whether the isomorphism problem for TSTC phylogenetic networks on a given set S of

taxa with globally bounded in-degree hybrid nodes (but without bounding the out-degree of the tree nodes; otherwise, Luks' theorem [8] would apply) is in P, but we conjecture that this is the case.

Acknowledgements

The authors would like to thank Antoni Lozano for his comments on an early version of this manuscript. The research described in this paper has been partially supported by the Spanish DGI project MTM2006-07773 COMGRIO.

References

[1] M. Baroni, C. Semple, M. Steel, Hybrids in real time. *Syst. Biol.* 55 (2006) 46–56.

[2] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente, A Distance Metric for a class of Tree-Sibling Phylogenetic Networks. *Bioinformatics* 24 (2008) 1481–1488.

[3] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente, Metrics for Phylogenetic Networks I: Generalizations of the Robinson-Foulds Metric. *IEEE T. Comput. Biol.* 6 (2009) 46–61.

[4] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente, Metrics for Phylogenetic Networks II: Nodal and Triplets Metric. *IEEE T. Comput. Biol.* (2009), in press.

[5] G. Cardona, F. Rosselló, G. Valiente, Tripartitions do not always discriminate phylogenetic networks. *Math. Biosci.* 211 (2008) 356–370.

[6] W. F. Doolittle, Phylogenetic classification and the universal tree. *Science* 284 (1999) 2124–2128.

[7] M. Goldberg. The Graph Isomorphism Problem. In: *Handbook of Graph Theory* (J. L. Gross and J. Yellen eds.) (CRC Press, 2003) 68–78.

[8] E. M. Luks, Isomorphism of Graphs of Bounded Valence can be Tested in Polynomial Time. *J. Comput. Syst. Sci.* 25 (1982) 42–65.

[9] G. Jin, L. Nakhleh, S. Snir, T. Tuller, Maximum likelihood of phylogenetic networks. *Bioinformatics* 22 (2006) 2604–2611

[10] G. Jin, L. Nakhleh, S. Snir, T. Tuller, Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics* 23 (2007) 123–128

[11] J. Köbler, U. Schöning, J. Torán, *The graph isomorphism problem: Its structural complexity*. Birkhäuser (1993).

[12] B. M. E. Moret, L. Nakhleh, T. Warnow, C. R. Linder, A. Tholse, A. Padolina, J. Sun, R. Timme, Phylogenetic networks: Modeling, reconstructibility, and accuracy. *IEEE T. Comput. Biol.* 1 (2004) 13–23.

[13] L. Nakhleh, *Phylogenetic networks*. PhD thesis, University of Texas at Austin (2004).

[14] L. Nakhleh, J. Sun, T. Warnow, C. R. Linder, B. M. E. Moret, A. Tholse, Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. *Proc. 8th Pacific Symp. Biocomputing* (2003) 315–326.

[15] V. N. Zemlyachenko, N. M. Korneenko, R. I. Tyshkevich, Graph isomorphism problem, *J. Math. Sci.* 29 (1985) 1426–1481.

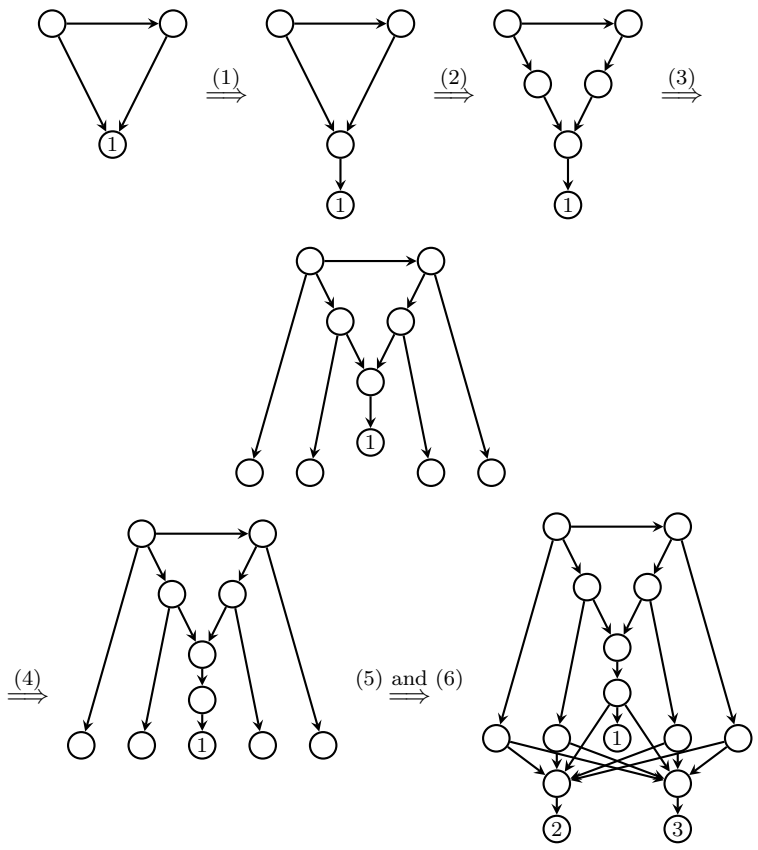

Figure 3: An example of the construction involved in the reduction of the isomorphism of S-rDAGs to the isomorphism of TSTC networks.